INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Variable selection in model-based clustering: A general variable role modeling

Cathy Maugis — Gilles Celeux — Marie-Laure Martin-Magniette

N° 6744

Novembre 2008

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6744--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Variable selection in model-based clustering: A general variable role modeling

Cathy Maugis [*] , Gilles Celeux [†] , Marie-Laure Martin-Magniette [‡§]

Thème COG — Systèmes cognitifs
Équipe-Projet SELECT

Rapport de recherche n° 6744 — Novembre 2008 — 21 pages

Abstract: The currently available variable selection procedures in model-based clustering assume that the irrelevant clustering variables are all independent or are all linked with the relevant clustering variables. We propose a more versatile variable selection model which describes three possible roles for each variable: The relevant clustering variables, the irrelevant clustering variables dependent on a part of the relevant clustering variables and the irrelevant clustering variables totally independent of all the relevant variables. A model selection criterion and a variable selection algorithm are derived for this new variable role modeling. The model identifiability and the consistency of the variable selection criterion are also established. Numerical experiments highlight the interest of this new modeling.

Key-words: Relevant, redundant or independent variables, Variable selection, Model-based clustering, Linear regression, BIC

[*] Université Paris-Sud 11,Projet SELECT
[†] INRIA Saclay - Île-de-France, Projet SELECT, Université Paris-Sud 11
[‡] UMR AgroParisTech/INRA MIA 518, Paris
[§] URGV UMR INRA 1165, CNRS 8114, UEVE, Evry

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Sélection de variables pour la classification non supervisée par mélanges gaussiens : une modélisation générale du rôle des variables

Résumé : Les procédures de sélection de variables actuellement disponibles en classification non supervisée par mélanges gaussiens supposent que les variables non significatives pour la classification sont toutes indépendantes ou sont toutes liées aux variables significatives. Nous proposons un modèle de sélection de variables plus général qui permet pour chaque variable d'être une variable significative pour la classification, d'être non significative mais dépendante d'une partie ou de toutes les variables significatives ou d'être non significative et indépendante des variables significatives. Le critère de sélection de modèles et l'algorithme de sélection de variables sont établis pour cette nouvelle modélisation. L'identifiabilité des modèles et la consistance du critère de sélection sont également établis. Des exemples numériques mettent en évidence l'intérêt de cette nouvelle modélisation.

Mots-clés : Variables significatives, redondantes ou indépendantes, Sélection de variables, Classification non supervisée, Mélanges gaussiens, Régression linéaire, BIC

1 Introduction

Model-based cluster analysis is making use of a mixture model to define subpopulations associated to mixture components. Since this approach is based on a probabilistic model, it provides a well-ground setting to answer important questions as the choice of a sensible number of clusters (McLachlan and Peel, 2000) and their relevance. Recently, several authors have considered that the structure of interest for the clustering may be contained in a subset of the available variables. They have recast the variable selection for clustering in the setting of Gaussian mixtures. We can cite among others the contributions of Law et al. (2004), Raftery and Dean (2006), Tadesse et al. (2005) and Maugis et al. (2008).

Among the variable selection procedures available for the clustering with Gaussian mixture models, the procedure of Law et al. (2004) assumed the irrelevant variables to be independent of the relevant clustering variables. Raftery and Dean (2006) proposed a first answer to this limitation by assuming that the irrelevant variables are regressed on the whole relevant variable set. Their modeling enforced the dependency link between the two types of variables. In Maugis et al. (2008), an improvement of Raftery and Dean's approach has been suggested by allowing the irrelevant variables to be explained by only a relevant variable subset. Models in competition are composed of the relevant clustering variables S, the subset R of S required to explain the irrelevant variables according to a linear regression, in addition with the number of mixture components K and the Gaussian mixture form m. The selected model is the maximizer of the BIC approximation of the integrated likelihood. These models cover in particular Raftery and Dean's modeling since R can be equal to S and also the approach of Law et al. (2004) since R can be an empty subset. Nevertheless, this variable selection model is not completely general since it does not allow some irrelevant variables to be independent and others to be dependent of the relevant variables simultaneously. For this very reason, it could imply an overpenalization of some models, in particular when the more parsimonious Gaussian mixture models are employed, as illustrated with the following example.

The dataset consists of $n = 2000$ data points $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ such that $\mathbf{y}_i' \in \mathbb{R}^{14}$. For the first two variables data are distributed from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (0,0)$, $\mu_2 = (4,0)$, $\mu_3 = (0,2)$ and $\mu_4 = (4,2)$. The dataset representation on the first two variables is given in Figure 1. The third variable is defined by $\mathbf{y}^3 = 0.5\mathbf{y}^1 + \mathbf{y}^2 + \varepsilon$, ε being sampled from a $\mathcal{N}(0, I_{2000})$ density. Finally eleven noisy variables are also appended: For each individual i, $\mathbf{y}_i^{\{4,\ldots,14\}}$ is simulated according to $\mathcal{N}((0, 0.4, 0.8, \ldots, 3.6, 4), I_{11})$. Using the variable selection procedure of Maugis et al. (2008), the selected model is

$$(\hat{K} = 2, \hat{m} = [pLI], \hat{S} = \{1, 5, 7, 10 - 12\}, \hat{R} = \{1\})$$

and not the true model ($K_0 = 4, m_0 = [pLI], S_0 = \{1,2\}, R_0 = \{1,2\}$). Indeed, there is a dilemma to choose a clustering with two or four clusters as it can be seen in Figure 1. The selected model has 46 free parameters while the true model yields to 123 free parameters because several regression coefficients equal to zero are assumed to be free. It implies a huge increase in the BIC penalty which cannot be compensated by an increase of the loglikelihood.

In order to remedy to this drawback, we propose to refine the variable role modeling and to take into account the possibility that some irrelevant clustering

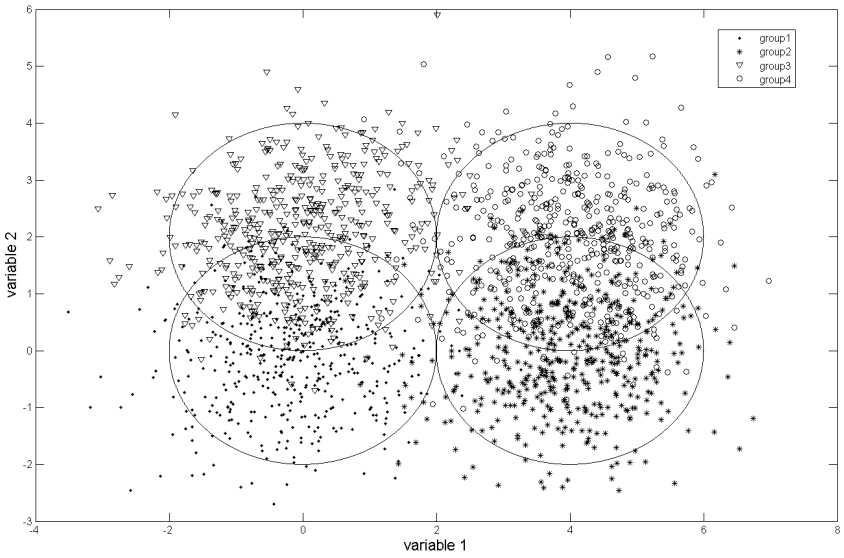

Figure 1: Representation of the dataset according to the first two variables.

variables are independent of all the relevant clustering variables and others are linked to some relevant variables at the same time. Such modeling allows us to define completely the variable role by precising the relevant variables for the clustering, the redundant variables defined as irrelevant variables linked to some relevant variables and, the independent variables defined as irrelevant variables independent of all the relevant variables. Acting in such a way, we hope to improve the dataset clustering and the variable role analysis. Moreover, we propose an algorithm taking the new variable role modeling into account without sensitive increase of the computing time compared to the algorithm of Maugis et al. (2008).

The paper is organized as follows. The model involving three different possible roles of the variables is presented in Section 2. Section 3 is devoted to the presentation of the variable selection criterion related to this model. The model identifiability and the consistency of the variable selection criterion are analyzed in Section 4. The new backward variable selection algorithm is described in Section 5 and experimented on several datasets in Section 6. Finally, a discussion on the overall method is given in Section 7.

2 The variable selection model

A sample of n individuals $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ described by Q variables is considered. Our aim is to determine subpopulations of these individuals using a Gaussian mixture model. When there are numerous variables, it can be sensible to choose which variables are entering in the mixture model since the structure of interest may often be contained in a subset of the available variables and a lot of variables may be useless or even harmful. It is thus important to select the relevant variables from the cluster analysis view point. Determining the variable role in the mixture

distribution can be regarded as a model selection problem as well. The variable selection model we propose is as follows: The nonempty set of relevant clustering variables is denoted S. Its complement S^c containing the irrelevant clustering variables is divided into two variable subsets U and W. The variables belonging to U are explained by a variable subset R of S according to a linear regression while the variables in W are assumed to be independent of all the relevant variables. Note that if U is empty, R is empty too and otherwise R is assumed to be no empty. Denoting $\mathcal{F}$ the family of variable index subsets of $\{1,\ldots,Q\}$, the variable partition set can be described as follows:

$$\mathcal{V} = \left\{ (S,R,U,W) \in \mathcal{F}^4; \begin{array}{l} S \cup U \cup W = \{1,\ldots,Q\} \\ S \cap U = \emptyset, S \cap W = \emptyset, U \cap W = \emptyset \\ S \neq \emptyset, R \subseteq S \\ R = \emptyset \text{ if } U = \emptyset \text{ and } R \neq \emptyset \text{ otherwise} \end{array} \right\}.$$

Throughout this paper, a quadruplet (S,R,U,W) of $\mathcal{V}$ is denoted $\mathbf{V} = (S,R,U,W)$. This new variable partition is summarized in Figure 2.

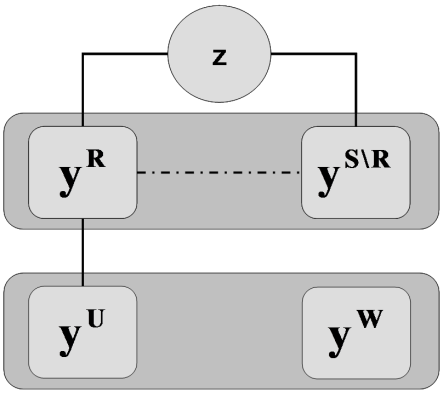

Figure 2: Graphical representation of a variable partition $\mathbf{V} = (S,R,U,W)$.

The density family associated to a variable partition $\mathbf{V}$ is decomposed into three subfamilies of densities related to the three possible variable roles and thus the unknown density h of the sample $\mathbf{y}$ is modeled by the product of three terms f_{clust}, f_{reg} and f_{indep} that are now specified. On the relevant variable subset S, a Gaussian mixture is considered. It is characterized by its number of clusters K and its form m, essentially related to the assumptions on the component variance matrices (see for instance Biernacki et al., 2006). The set of such models (K,m) is denoted $\mathcal{T}$ and the likelihood on S for a given (K,m) is

$$f_{\text{clust}}(\mathbf{y}^S|K,m,\alpha) = \sum_{k=1}^{K} p_k \Phi(\mathbf{y}^S|\mu_k,\Sigma_k)$$

where the parameter vector is $\alpha = (p_1,\ldots,p_K,\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K)$, with $\sum_{k=1}^{K} p_k = 1$, the proportion vector and the variance matrices satisfying the form m.

The variables of the subset U are explained by the variables of the subset R according to a multidimensional linear regression where the variance matrix can

be assumed to have a spherical, diagonal or general form. These forms are denoted $[LI]$, $[LB]$ and $[LC]$ respectively by analogy with the notation of Gaussian mixture models (see Biernacki et al., 2006). The variance matrix form is thus specified by $r \in \mathcal{T}_{\text{reg}} = \{[LI], [LB], [LC]\}$. The likelihood associated to the linear regression of $\mathbf{y}^U$ on $\mathbf{y}^R$ is then

$$f_{\text{reg}}(\mathbf{y}^U | r, a + \mathbf{y}^R\beta, \Omega) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^U | a + \mathbf{y}_i^R\beta, \Omega)$$

where a is the $1 \times \text{Card}(U)$ intercept vector, β is the $\text{Card}(R) \times \text{Card}(U)$ coefficient regression matrix and Ω is the $\text{Card}(U) \times \text{Card}(U)$ variance matrix.

The marginal distribution of the data on the variable subset W, which contains the variables independent of all relevant variables, is assumed to be a Gaussian distribution with mean vector γ and variance matrix τ. The form of the variance matrix τ can be spherical or diagonal and is specified by $l \in \mathcal{T}_{\text{indep}} = \{[LI], [LB]\}$. The associated likelihood on W is then

$$f_{\text{indep}}(\mathbf{y}^W | l, \gamma, \tau) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^W | \gamma, \tau).$$

Finally, the model family is

$$\mathcal{N} = \{(K, m, r, l, \mathbf{V}); (K, m) \in \mathcal{T}, r \in \mathcal{T}_{\text{reg}}, l \in \mathcal{T}_{\text{indep}}, \mathbf{V} \in \mathcal{V}\} \tag{1}$$

and the likelihood for a model $(K, m, r, l, \mathbf{V})$ is given by

$$f(\mathbf{y}|K, m, r, l, \mathbf{V}, \theta) = f_{\text{clust}}(\mathbf{y}^S | K, m, \alpha) f_{\text{reg}}(\mathbf{y}^U | r, a + \mathbf{y}^R\beta, \Omega) f_{\text{indep}}(\mathbf{y}^W | l, \gamma, \tau)$$

where the parameter vector $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$ belongs to a parameter vector set $\Upsilon_{(K,m,r,l,\mathbf{V})}$.

This model based on a quadruplet (S, R, U, W) involves the three possible variable roles. In the following, it is called SRUW. It is a generalization of the model proposed in Maugis et al. (2008), associated to a couple (S, R), since it can be interpreted as a SRUW model with $U = S^c$ and $W = \emptyset$. In what follows, this previous model will be referred as SR model.

3 Model selection criterion

The new model collection SRUW allows to recast the variable selection problem for clustering into a model selection problem. Ideally, we search the model maximizing the integrated loglikelihood

$$(\tilde{K}, \tilde{m}, \tilde{r}, \tilde{l}, \tilde{\mathbf{V}}) = \underset{(K,m,r,l,\mathbf{V}) \in \mathcal{N}}{\text{argmax}} \ln\{f(\mathbf{y}|K, m, r, l, \mathbf{V})\}$$

where the integrated likelihood can be decomposed into

$$f(\mathbf{y}|K, m, r, l, \mathbf{V}) = f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{reg}}(\mathbf{y}^U | r, \mathbf{y}^R) f_{\text{indep}}(\mathbf{y}^W | l) \tag{2}$$

with

$$f_{\text{clust}}(\mathbf{y}^S | K, m) = \int f_{\text{clust}}(\mathbf{y}^S | K, m, \alpha) \pi(\alpha | K, m) d\alpha,$$

$$f_{\text{reg}}(\mathbf{y}^U|r,\mathbf{y}^R) = \int f_{\text{reg}}(\mathbf{y}^U|r,a+\mathbf{y}^R\beta,\Omega)\pi(a,\beta,\Omega|r)d(a,\beta,\Omega)$$

and

$$f_{\text{indep}}(\mathbf{y}^W|l) = \int f_{\text{indep}}(\mathbf{y}^W|l,\gamma,\tau)\pi(\gamma,\tau|l)d(\gamma,\tau).$$

The three functions π are the prior distributions of the different vector parameters. Since these integrated likelihoods are difficult to evaluate, they are approximated by their associated BIC criterion.

Bayesian Information Criterion for Gaussian mixture The BIC criterion associated to the Gaussian mixture on the relevant variable subset S is given by

$$\text{BIC}_{\text{clust}}(\mathbf{y}^S|K,m) = 2\ln[f_{\text{clust}}(\mathbf{y}^S|K,m,\hat{\alpha})] - \lambda_{(K,m,S)}\ln(n) \tag{3}$$

where $\hat{\alpha}$ is the maximum likelihood estimator, obtained using the EM algorithm (Dempster et al., 1977), and $\lambda_{(K,m,S)}$ is the number of free parameters of this Gaussian mixture model (K,m) on the variable subset S (see Biernacki et al., 2006).

Bayesian Information Criterion for linear regression For the linear regression of the variable subset U on R, the associated BIC criterion is defined by

$$\text{BIC}_{\text{reg}}(\mathbf{y}^U|r,\mathbf{y}^R) = 2\ln[f_{\text{reg}}(\mathbf{y}^U|r,\hat{a}+\mathbf{y}^R\hat{\beta},\hat{\Omega})] - \nu_{(r,U,R)}\ln(n) \tag{4}$$

where $\hat{a}$, $\hat{\beta}$ and $\hat{\Omega}$ are the maximum likelihood estimators. The estimated intercept vector and the regression coefficient matrix are given by $(\hat{a}',\hat{\beta}')' = (X'X)^{-1}X'\mathbf{y}^U$ where $X = (1_n,\mathbf{y}^R)$, 1_n being a n-vector of ones. The estimated variance matrix $\hat{\Omega}$ and the number of free parameters in this linear regression denoted $\nu_{(r,U,R)}$ depend on the form index r. If r assigns the general form ($r = [LC]$), the estimated variance matrix is given by

$$\hat{\Omega} = \frac{1}{n}\mathbf{y}^{U\,\prime}\{I_n - X(X'X)^{-1}X'\}\mathbf{y}^U$$

and the number of free parameters is equal to

$$\nu_{(r,U,R)} = \text{Card}(U)\times\{\text{Card}(R)+1\} + \frac{\text{Card}(U)\{\text{Card}(U)+1\}}{2}.$$

If r is the diagonal form ($r = [LB]$), the estimated variance matrix is written $\hat{\Omega} = \text{diag}(\hat{\omega}_1^2,\ldots,\hat{\omega}_{\text{Card}(U)}^2)$ where the diagonal elements are defined by

$$\hat{\omega}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^U - \hat{a} - \mathbf{y}_i^R\hat{\beta})_j^2,\ \forall j\in\{1,\ldots,\text{Card}(U)\}$$

and the number of free parameters is $\nu_{(r,U,R)} = \text{Card}(U)\times\{\text{Card}(R)+1\}+\text{Card}(U)$. When r assigns the spherical form ($r = [LI]$), the estimated variance matrix is equal to $\hat{\Omega} = \hat{\omega}^2 I_{\text{Card}(U)}$ where

$$\hat{\omega}^2 = \frac{1}{n\,\text{Card}(U)}\sum_{i=1}^{n}\|\mathbf{y}_i^U - \hat{a} - \mathbf{y}_i^R\hat{\beta}\|^2,$$

$\|.\|$ denoting the l_2 norm, and the number of free parameters is $\nu_{(r,U,R)} = \text{Card}(U)\times\{\text{Card}(R)+1\}+1$.

Bayesian Information Criterion for a Gaussian density The BIC criterion associated to the Gaussian density on the variable subset W is given by

$$\mathrm{BIC}_{\mathrm{indep}}(\mathbf{y}^W|l) = 2\ln[f_{\mathrm{indep}}(\mathbf{y}^W|l,\hat{\gamma},\hat{\tau})] - \rho_{(l,W)}\ln(n). \tag{5}$$

The parameters $\hat{\gamma}$ and $\hat{\tau}$ denote the maximum likelihood estimators and $\rho_{(l,W)}$ is the number of free parameters. Whatever the form of the variance matrices, the estimated mean vector is given by

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^W.$$

If l assigns the diagonal form ($l = [LB]$), the estimated variance matrix is expressed as $\hat{\tau} = \mathrm{diag}(\hat{\sigma}_1^2,\ldots,\hat{\sigma}_{\mathrm{Card}(W)}^2)$ where the diagonal elements are given by

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^W - \hat{\gamma})_j^2,\ \forall j \in \{1,\ldots,\mathrm{Card}(W)\}$$

and the number of free parameters is equal to $\rho_{(l,W)} = 2\,\mathrm{Card}(W)$. Otherwise, l indicating the spherical form ($l = [LI]$), the estimated variance matrix is $\hat{\tau} = \hat{\sigma}^2 I_{\mathrm{Card}(W)}$ where

$$\hat{\sigma}^2 = \frac{1}{n\,\mathrm{Card}(W)}\sum_{i=1}^{n}\|\mathbf{y}_i^W - \hat{\gamma}\|^2$$

and the number of free parameters is equal to $\rho_{(l,W)} = \mathrm{Card}(W) + 1$.

Finally, the three terms of the integrated likelihood (2) are replaced with their BIC approximations (3), (4) and (5) respectively. Then the selected model satisfies

$$(\hat{K},\hat{m},\hat{r},\hat{l},\hat{\mathbf{V}}) = \underset{(K,m,r,l,\mathbf{V})\in\mathcal{N}}{\mathrm{argmax}}\ \mathrm{crit}(K,m,r,l,\mathbf{V}) \tag{6}$$

where the model selection criterion is the sum of the three BIC criteria

$$\mathrm{crit}(K,m,r,l,\mathbf{V}) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K,m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^U|r,\mathbf{y}^R) + \mathrm{BIC}_{\mathrm{indep}}(\mathbf{y}^W|l).$$

This criterion can also be written

$$\mathrm{crit}(K,m,r,l,\mathbf{V}) = 2\ln[f(\mathbf{y}|K,m,r,l,\mathbf{V},\hat{\theta})] - \Xi_{(K,m,r,l,\mathbf{V})}\ln(n) \tag{7}$$

where the maximum likelihood estimator is $\hat{\theta} = (\hat{\alpha},\hat{a},\hat{\beta},\hat{\Omega},\hat{\gamma},\hat{\tau})$ and the overall number of free parameters is $\Xi_{(K,m,r,l,\mathbf{V})} = \lambda_{(K,m,S)} + \nu_{(r,U,R)} + \rho_{(l,W)}$.

4 Theoretical properties

The theoretical properties established in Maugis et al. (2008) for SR model are generalized to SRUW model. First, necessary and sufficient conditions are given to ensure the identifiability of the SRUW model collections. Second, a consistency theorem of our variable selection criterion is stated.

4.1 Identifiability

The identifiability characterization is based on the identifiability of SR model collection and the difference between the variables in U and W.

Theorem 1. *Let $\Theta_{(K,m,r,l,\mathbf{V})}$ be a subset of the parameter set $\Upsilon_{(K,m,r,l,\mathbf{V})}$ such that elements $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$*

- *contain distinct couples (μ_k, Σ_k) fulfilling $\forall s \subsetneq S, \exists (k,k'), 1 \leq k < k' \leq K$;*

$$\mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s}, \tag{8}$$

  where $\bar{s}$ denotes the complement in S of any nonempty subset s of S

- *if $U \neq \emptyset$,*
  - *for all variables j of R, there exists a variable u of U such that the restriction β_{uj} of the regression coefficient matrix β associated to j and u is not equal to zero.*
  - *for all variables u of U, there exists a variable j of R such that $\beta_{uj} \neq 0$.*
- *parameters Ω and τ exactly respect the forms r and l respectively: They are both diagonal matrices with at least two different eigenvalues if $r = [LB]$ and $l = [LB]$ and Ω has at least a non-zero entry outside the main diagonal if $r = [LC]$.*

Let $(K, m, r, l, \mathbf{V})$ and $(K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ be two models. If there exist $\theta \in \Theta_{(K,m,r,l,\mathbf{V})}$ and $\theta^\star \in \Theta_{(K^\star,m^\star,r^\star,l^\star,\mathbf{V}^\star)}$ such that

$$f(.|K, m, r, l, \mathbf{V}, \theta) = f(.|K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star)$$

then $(K, m, r, l, \mathbf{V}) = (K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ and $\theta = \theta^\star$ (up to a permutation of mixture components).

Proof. This proof is based on the identifiability of SR models stated in Maugis et al. (2008) and proved in the associated Web Supplementary Materials or in Maugis (2008). First, we remark that for all row vector x of size Q,

$$\begin{aligned} f_{\text{reg}}(x^U|r, a + x^R\beta, \Omega) f_{\text{indep}}(x^W|l, \gamma, \tau) &= \Phi(x^U|a + x^R\beta, \Omega)\Phi(x^W|\gamma, \tau) \\ &= \Phi(x^{U\cup W}|\tilde{a} + x^R\tilde{\beta}, \tilde{\Omega}) \end{aligned}$$

where $\tilde{a} = (a, \gamma)$, $\tilde{\beta} = (\beta, 0)$ and $\tilde{\Omega}$ is the block diagonal matrix with diagonal elements Ω and τ. This remark allows us to consider parameter vectors $\tilde{\theta} = (\alpha, \tilde{a}, \tilde{\beta}, \tilde{\Omega})$ in the model (K, m, S, R) (among the SR model collection) in order to rewrite the densities in the following way

$$f(x|K, m, r, l, \mathbf{V}, \theta) = \tilde{f}_{\text{clust}}(x^S|K, m, \alpha)\tilde{f}_{\text{reg}}(x^{S^c}|\tilde{a} + x^R\tilde{\beta}, \tilde{\Omega}) = \tilde{f}(x|K, m, S, R, \tilde{\theta})$$

where $\tilde{f}_{\text{clust}}$, $\tilde{f}_{\text{reg}}$ and $\tilde{f}$ denote the density functions used under the SR modeling in Maugis et al. (2008). In the same way, $f(.|K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star) = \tilde{f}(.|K^\star, m^\star, S^\star, R^\star, \tilde{\theta}^\star)$. According to Hypothesis (8) and the identifiability property for the SR modeling, the equality

$$\tilde{f}_{\text{clust}}(x^S|K, m, \alpha)\tilde{f}_{\text{reg}}(x^{S^c}|\tilde{a}+x^R\tilde{\beta}, \tilde{\Omega}) = \tilde{f}_{\text{clust}}(x^{S^\star}|K^\star, m^\star, \alpha^\star)\tilde{f}_{\text{reg}}(x^{S^{\star\,c}}|\tilde{a}^\star+x^{R^\star}\tilde{\beta}^\star, \tilde{\Omega}^\star)$$

implies that $K = K^\star$, $m = m^\star$, $\alpha = \alpha^\star$, $S = S^\star$, $R = R^\star$, $\tilde{a} = \tilde{a}^\star$, $\tilde{\beta} = \tilde{\beta}^\star$ and $\tilde{\Omega} = \tilde{\Omega}^\star$. Then we consider the decompositions $S^c = U \cup W$ and $S^{\star\, c} = U^\star \cup W^\star$ knowing that $S^c = S^{\star\, c}$. If there exists a variable j belonging to $U^\star \cap W$ then for all $q \in R, (\beta, 0)_{qj} = 0 = (\beta^\star, 0)_{qj}$ and there exists $q \in R^\star = R$ such that $\beta^\star_{qj} \neq 0$. Thus by contradiction, we obtain that $U^\star \cap W$ is empty and in the same way, $U \cap W^\star$ is an empty set. Finally, it leads to $W = W^\star$, $U = U^\star$ and, identifying each parameter term $\tilde{a}$, $\tilde{\beta}$ and $\tilde{\Omega}$, we obtain that $a = a^\star$, $\beta = \beta^\star$, $\gamma = \gamma^\star$, $\tau = \tau^\star$, $\Omega = \Omega^\star$ and then $r = r^\star$ and $l = l^\star$. □

4.2 Consistency of our criterion

As for SR model, a consistency property of the criterion restricted to the variable partition selection for SRUW model can be achieved. In this section, it is proved that the probability of selecting the true variable partition $\mathbf{V}_0 = (S_0, R_0, U_0, W_0)$ by maximizing Criterion (7) approaches 1 as $n \to \infty$ when the sampling distribution is one of the densities in competition and the true model (K_0, m_0, r_0, l_0) is known. Denoting h the density function of the sample $\mathbf{y}$, the two following vectors are considered

$$\begin{aligned}\theta^\star_{(K,m,r,l,\mathbf{V})} &= \underset{\theta_{(K,m,r,l,\mathbf{V})} \in \Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmin}} \mathrm{KL}[h, f(.|\theta_{(K,m,r,l,\mathbf{V})})] \\ &= \underset{\theta_{(K,m,r,l,\mathbf{V})} \in \Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmax}} \mathbb{E}_X\{\ln f(X|\theta_{(K,m,r,l,\mathbf{V})})\},\end{aligned}$$

where $\mathrm{KL}[h, f] = \int \ln\left\{\frac{h(x)}{f(x)}\right\} h(x)dx$ is the Kullback-Leibler divergence between the densities h and f and

$$\hat{\theta}_{(K,m,r,l,\mathbf{V})} = \underset{\theta_{(K,m,r,l,\mathbf{V})} \in \Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln\{f(\mathbf{y}_i|\theta_{(K,m,r,l,\mathbf{V})})\}.$$

Recall that $\Theta_{(K,m,r,l,\mathbf{V})}$ is the subset defined in Theorem 1 where the model identifiability is ensured.

The following assumption is considered:

(H1) The density h is assumed to be one of the densities in competition. By identifiability, there exists a unique model $(K_0, m_0, r_0, l_0, \mathbf{V}_0)$ and an associated parameter $\theta^\star_{(K_0,\, m_0,\, r_0,\, l_0,\, \mathbf{V}_0)}$ such that $h = f(.|\theta^\star_{(K_0,\, m_0,\, r_0,\, l_0,\, \mathbf{V}_0)})$. The model (K_0, m_0, r_0, l_0) is supposed to be known.

To simplify the notation, all the dependencies over this model (K_0, m_0, r_0, l_0) are omitted in the following. Moreover, an additional technical assumption is considered:

(H2) The vectors $\theta^\star_{\mathbf{V}}$ and $\hat{\theta}_{\mathbf{V}}$ are supposed to belong to a compact subspace $\Theta'_{\mathbf{V}}$ of the following subset

$$\left(\begin{array}{c}\mathcal{P}_{K-1} \times \mathcal{B}(\eta, \mathrm{card}(S))^{K_0} \times \mathcal{D}^{K_0}_{\mathrm{card}(S)} \times \mathcal{B}(\rho, \mathrm{card}(U)) \\ \times \mathcal{B}(\rho, \mathrm{card}(R), \mathrm{card}(U)) \times \mathcal{D}_{\mathrm{card}(U)} \times \mathcal{B}(\eta_1, \mathrm{card}(W)) \times \mathcal{D}_{\mathrm{card}(W)}\end{array}\right) \cap \Theta_{\mathbf{V}}$$

where

- $\mathcal{P}_{K-1} = \left\{(p_1, \ldots, p_K) \in [0,1]^K; \sum_{k=1}^{K} p_k = 1\right\}$ denotes the $K-1$ dimensional simplex containing the considered proportion vectors,
- $\mathcal{B}(\eta, r)$ is the closed ball in $\mathbb{R}^r$ of radius η centered at zero for the l^2-norm defined by $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{r} x_i^2}, \forall \mathbf{x} \in \mathbb{R}^r$,
- $\mathcal{B}(\rho, r, q)$ is the closed ball in $\mathcal{M}_{r\times q}(\mathbb{R})$ of radius ρ centered at zero for the matricial norm $|||.|||$ defined by

$$\forall A \in \mathcal{M}_{r\times q}(\mathbb{R}), |||A||| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{x}A\|,$$

- $\mathcal{D}_r$ is the set of the $r \times r$ positive definite matrices with eigenvalues in $[s_{\mathrm{m}}, s_{\mathrm{M}}]$ with $0 < s_{\mathrm{m}} < s_{\mathrm{M}}$.

Theorem 2. *Under assumptions (H1) and (H2), the variable partition* $\hat{\mathbf{V}} = (\hat{S}, \hat{R}, \hat{U}, \hat{W})$ *maximizing Criterion (7) with fixed* (K_0, m_0, r_0, l_0) *is such that*

$$P(\hat{\mathbf{V}} = \mathbf{V}_0) = P((\hat{S}, \hat{R}, \hat{U}, \hat{W}) = (S_0, R_0, U_0, W_0)) \underset{n\to\infty}{\to} 1.$$

The theorem proof is given in Appendix A.

5 The variable selection procedure

An exhaustive research of the model maximizing Criterion (7) is impossible since the number of models is huge. Thus we design a procedure, embedding backward stepwise algorithms to determine the best variable roles.

5.1 The models in competition

At a fixed step of the algorithm, the variable set $\{1, \ldots, Q\}$ is divided into the set of selected clustering variables S, the set U of irrelevant variables which are linked to some relevant variables, the set W of independent irrelevant variables and j the candidate variable for inclusion into or exclusion from the clustering variable set. Under the model (K, m, r, l), the integrated likelihood can be decomposed as

$$\begin{aligned} f(\mathbf{y}^S, \mathbf{y}^j, \mathbf{y}^U, \mathbf{y}^W | K, m, r, l) &= f(\mathbf{y}^U, \mathbf{y}^W | \mathbf{y}^S, \mathbf{y}^j, K, m, r, l) f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) \\ &= f_{\mathrm{indep}}(\mathbf{y}^W | l) f_{\mathrm{reg}}(\mathbf{y}^U | r, \mathbf{y}^S, \mathbf{y}^j) f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l). \end{aligned}$$

Three situations can then occur for the candidate variable j:

- M1: Given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering,

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m).$$

- M2: Given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering but has a linear link with the variables of $R[j]$ (the nonempty subset of S containing the relevant variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^S$),

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\mathrm{clust}}(\mathbf{y}^S | K, m) f_{\mathrm{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{R[j]}).$$

- M3: Given $\mathbf{y}^S$, $\mathbf{y}^j$ is independent of all the variables of S,

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{indep}}(\mathbf{y}^j | [LI]).$$

In models M2 and M3, the form of the variance matrices in the regression and in the Gaussian density are $r = [LI]$ and $l = [LI]$ respectively since j is a single variable. In order to compare these three situations in an efficient way, we remark that $f_{\text{indep}}(\mathbf{y}^j | [LI])$ can be written $f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^\emptyset)$. Thus instead of considering the nonempty subset $R[j]$ we consider a new explicative variable subset denoted $\tilde{R}[j]$ and defined by $\tilde{R}[j] = \emptyset$ if j follows model M3 and $\tilde{R}[j] = R[j]$ if j follows model M2. It allows us to recast the comparison of the three models into the comparison of two models with the Bayes factor

$$\frac{f_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m)}{f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\tilde{R}[j]})}.$$

This Bayes factor being difficult to evaluate, it is approximated by

$$\text{BIC}_{\text{diff}}(j) = \text{BIC}_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m) - \left\{ \text{BIC}_{\text{clust}}(\mathbf{y}^S | K, m) + \text{BIC}_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\tilde{R}[j]}) \right\}.$$

It is worth noticing that $\text{BIC}_{\text{diff}}(.)$ is the criterion used to construct the backward variable selection algorithm for SR model in Maugis et al. (2008).

5.2 The general steps of the algorithm

This algorithm makes use of the clustering variable selection backward algorithm and the regression backward variable selection algorithm for SR model.

▶ For each mixture model (K, m):

- The variable partition into $\hat{S}(K, m)$ and $\hat{S}^c(K, m)$ is determined by the backward stepwise selection algorithm described in Maugis et al. (2008).
- The variable subset $\hat{S}^c(K, m)$ is divided into $\hat{U}(K, m)$ and $\hat{W}(K, m)$: For each variable j belonging to $\hat{S}^c(K, m)$, the variable subset $\tilde{R}[j]$ of $\hat{S}(K, m)$ allowing to explain j by a linear regression is determined with the backward stepwise regression algorithm. If $\tilde{R}[j] = \emptyset$, $j \in \hat{W}(K, m)$ and otherwise, $j \in \hat{U}(K, m)$.
- For each form r:
  - ∗ The variable subset $\hat{R}(K, m, r)$, included into $\hat{S}(K, m)$ and explaining the variables of $\hat{U}(K, m)$, is determined using a backward stepwise regression algorithm with the fixed form regression model r.
  - ∗ For each form l: $\hat{\theta}$ and the following criterion value are computed

$$\widetilde{\text{crit}}(K, m, r, l) = \text{crit}(K, m, r, l, \hat{S}(K, m), \hat{R}(K, m, r), \hat{U}(K, m), \hat{W}(K, m)).$$

▶ The model satisfying the following condition is then selected

$$(\hat{K}, \hat{m}, \hat{r}, \hat{l}\,) = \underset{(K,m,r,l) \in \mathcal{T} \times \mathcal{T}_{\text{reg}} \times \mathcal{T}_{\text{indep}}}{\text{argmax}} \widetilde{\text{crit}}(K, m, r, l).$$

▶ Finally, the complete selected model is

$$\left(\hat{K}, \hat{m}, \hat{r}, \hat{l}, \hat{S}(\hat{K}, \hat{m}), \hat{R}(\hat{K}, \hat{m}, \hat{r}), \hat{U}(\hat{K}, \hat{m}), \hat{W}(\hat{K}, \hat{m})\right).$$

Remark: It is worth noticing that the complexity of the algorithm is not increased compared with the algorithm involved for SR model despite the three possible variable roles. It is due to the use of $\tilde{R}[j]$ defined in Section 5.1.

6 Method validation

This section is devoted to illustrate the behaviour of the SRUW variable selection method and compare it to the SR selection method. First, we study a simulated example where different scenarii for the irrelevant clustering variables are considered. In particular, this example contains the dataset considered in the introduction. Second, the study of the waveform dataset (see Breiman et al., 1984), which is not distributed from a Gaussian mixture, is performed.

6.1 Seven simulated situations

The dataset consists of 2000 data points in $\mathbb{R}^{14}$. For the first two variables data are distributed from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (0,0)$, $\mu_2 = (4,0)$, $\mu_3 = (0,2)$ and $\mu_4 = (4,2)$. The dataset representation, given in Figure 1, shows the difficulty to choose between 2 or 4 clusters for this dataset. Twelve variables have been appended, simulated according to $\mathbf{y}_i^{\{3,\dots,14\}} = \tilde{a} + \mathbf{y}_i^{\{1,2\}}\tilde{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \tilde{\Omega})$ and $\tilde{a} = (0, 0, 0.4, 0.8, \dots, 3.6, 4)$. The parameters $\tilde{\beta}$ and $\tilde{\Omega}$ have been chosen according to different scenarii ranging from all variables are independent of the relevant clustering variables to all irrelevant clustering variables depend on relevant clustering variables and with different forms for the variance matrices in the regression and the independent Gaussian density. These different scenarii are described in Table 1.

Scenario	$\tilde{\beta}$	$\tilde{\Omega}$
n° 1	0_{12}	I_{12}
n° 2	$((3,0)', 0_{11})$	$\mathrm{diag}(0.5, I_{11})$
n° 3	$((0.5,1)', 0_{11})$	I_{12}
n° 4	$(\beta_1, 0_{10})$	I_{12}
n° 5	$(\beta_1, \beta_2, 0_7)$	$\mathrm{diag}(I_3, 0.5I_5, I_4)$
n° 6	$(\beta_1, \beta_2, \beta_3, 0_3)$	$\mathrm{diag}(I_3, 0.5I_2, \Omega_1, \Omega_2, I_3)$
n° 7	$(\beta_1, \beta_2, \beta_3, (-1,-2)', (0,0.5)', (1,1)')$	$\mathrm{diag}(I_3, 0.5I_2, \Omega_1, \Omega_2, I_3)$

Table 1: Description of the seven scenarii where 0_p is the $2 \times p$ zero matrix, $\beta_1 = ((0.5,1)', (2,0)')$, $\beta_2 = ((0,3)', (-1,2)', (2,-4)')$, $\beta_3 = ((0.5,0)', (4,0.5)', (3,0)', (2,1)')$, $\Omega_1 = \mathrm{Rot}(\pi/3)'\,\mathrm{diag}(1,3) * \mathrm{Rot}(\pi/3)$, $\Omega_2 = \mathrm{Rot}(\pi/6)'\,\mathrm{diag}(2,6)\mathrm{Rot}(\pi/6)$, $\mathrm{Rot}(\theta)$ denoting the plane rotation matrix with angle θ.

The algorithms associated to SRUW and SR variable selection models are compared on these seven scenarii (see Tables 2 and 3). SR variable selection procedure has difficulties with the first six scenarii. It selects a spherical Gaussian

Scenario	$\hat{K}$	$\hat{m}$	$\hat{S}$	$\hat{R}$
n° 1	2	$[pLI]$	$\{1, 6, 8, 9, 12-14\}$	$\emptyset$
n° 2	2	$[pLI]$	$\{1, 4, 6\}$	$\{1\}$
n° 3	2	$[pLI]$	$\{1, 5, 7, 10-12\}$	$\{1\}$
n° 4	2	$[pLI]$	$\{1, 5-8, 11, 13\}$	$\{1\}$
n° 5	2	$[pLI]$	$\{1, 4\}$	$\{1\}$
n° 6	2	$[pLI]$	$\{1, 13, 14\}$	$\{1\}$
n° 7	4	$[pLC]$	$\{1, 2\}$	$\{1, 2\}$

Table 2: Model selection results obtained with SR variable selection method. The true model is composed of $K_0 = 4$, $m_0 = [pLI]$, $S_0 = \{1, 2\}$ and $R_0 = \emptyset$ for Scenario 1, $R_0 = \{1\}$ for Scenario 2 and $R_0 = \{1, 2\}$ for the other scenarii, with SR model.

Scenario	$\hat{K}$	$\hat{m}$	$\hat{r}$	$\hat{l}$	$\hat{S}$	$\hat{R}$	$\hat{U}$	$\hat{W}$
n° 1	4	$[pLI]$	-	$[LI]$	$\{1, 2\}$	$\emptyset$	$\emptyset$	$\{3-14\}$
n° 2	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1\}$	$\{3\}$	$\{4-14\}$
n° 3	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3\}$	$\{4-14\}$
n° 4	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3, 4\}$	$\{5-14\}$
n° 5	4	$[pLI]$	$[LB]$	$[LB]$	$\{1, 2\}$	$\{1, 2\}$	$\{3-7\}$	$\{8-14\}$
n° 6	4	$[pLI]$	$[LC]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3-11\}$	$\{12-14\}$
n° 7	4	$[pLC]$	$[LC]$	-	$\{1, 2\}$	$\{1, 2\}$	$\{3-14\}$	$\emptyset$

Table 3: Model selection results obtained with SRUW variable selection method. For all scenarii, the three first elements of the true model are $K_0 = 4$, $m_0 = [pLI]$ and $S_0 = \{1, 2\}$. The selected $\hat{r}$, $\hat{l}$, $\hat{R}$, $\hat{U}$ and $\hat{W}$ correspond to the true model elements for all scenarii.

mixture with two components. Although Variable 1 is the more significant and seems to be required alone to obtain such a clustering in two groups, the procedure selects besides some noise variables (see Table 2). SR method only succeeds in finding the true variable partition for Scenario 7 where irrelevant variables are all dependent to the relevant variables. The true number of clusters is well chosen for this dataset, but SR method selects a more complex Gaussian mixture form $[pLC]$. With the SRUW variable selection procedure, these difficulties of selection disappear. This new method selects the true variable partition and chooses a clustering in four clusters (see Table 3). The form of variance matrices for the regression and for the independent Gaussian density are correctly identified except for Scenario 7. This variable selection improvement is due to the use of a larger and more realistic model family and leads to a fairer penalization of the models. For instance in Scenario 3, the true distribution involves 123 parameters in SR model and 25 parameters in SRUW model.

6.2 Waveform dataset

This dataset is composed of 5000 points based on a random convex combination of two of three waveforms (see Fig 3) sampled at integers $\{1, \ldots, 21\}$ with noise added and nineteen noisy standard centered Gaussian variables are appended. A detailed description of the waveform dataset is available in Breiman et al. (1984). For SRUW method the number of components K belongs to $\{3, 4, 5, 6\}$ and twenty mixture forms are used (spherical forms, diagonal forms and the general forms assigned by $[p\_L\_C\_]$). It selects the Gaussian mixture model ($\hat{K} = 6, \hat{m} = [p_k LC]$) and a spherical form for the variance matrix in the regression and in the independent Gaussian density ($\hat{r} = [LI]$ and $\hat{l} = [LI]$), with the following variable partition

$$(\hat{S} = \{4{-}18\}, \hat{R} = \{5{-}7, 9{-}12, 14, 15, 17\}, \hat{U} = \{2, 3, 19, 20, 38\}, \hat{W} = \{1, 21{-}37, 39, 40\}).$$

SRUW method allows us to highlight that several variables are independent of the relevant variables. Except Variable 38, the standard centered Gaussian variables are declared independent. Moreover, it reveals that the link between the variables of $\hat{U}$ with the relevant variables is more complex. SR method selects the model ($\hat{K} = 6, \hat{m} = [p_k LC], \hat{S} = \{4 - 18\}, \hat{R} = \{7, 11, 15\}$). Only the maximum of each wave $\{7, 11, 15\}$ are selected to explain irrelevant variables because all the noise variables are regressed. With SRUW model, the independent variables being identified, analyzing the dependence of the irrelevant variables of $\hat{U}$ requires several relevant variables. It is more realistic since the dataset is based on a random convex combination of two of three waveforms (see Fig 3).

7 Discussion

A new modeling of the variable role in a model-based clustering setting has been proposed to improve the clustering and its interpretation. A large model family is considered to lead to a general variable selection model. In particular our model is relevant when the clustering is difficult to determine or when it is supported by spherical or diagonal Gaussian mixtures for which the variable selection is a sensitive task. Our SRUW model is versatile since it recovers all the possible variable

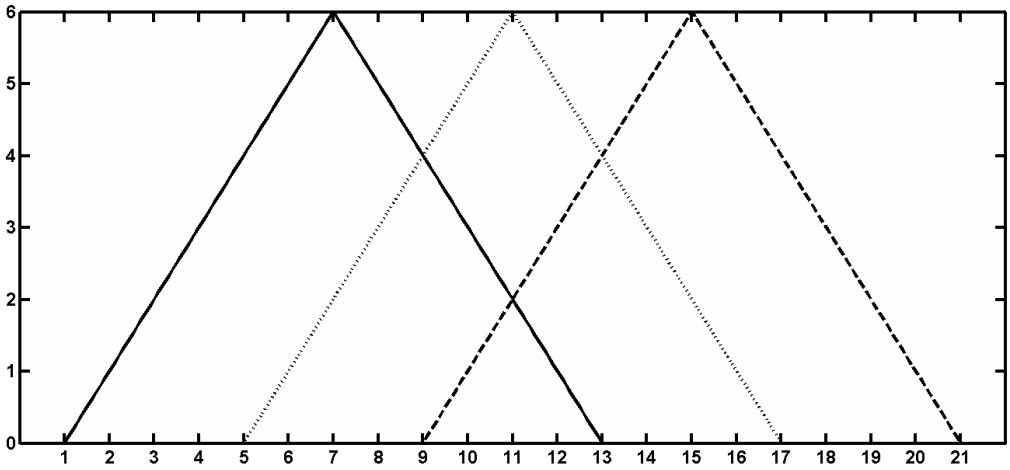

Figure 3: Representation of the three wave functions used to construct the waveform dataset.

roles: Significant (S), redundant (U in relation with R) and noisy (W). All previously studied models can be obtained as particular SRUW models. For instance, it can happen that $W = \emptyset$. It means that no independent variables are present. Thus in transcriptome examples as the one studied in Maugis et al. (2008), the selected variables under SR and SRUW models are often identical. From a biological point of view, this result gives an additional useful information since it highlights the complex relations between all transcriptomes. Theoretically, the model identifiability and the criterion consistency are extended to this more versatile model collection. Despite the richness of the model collection, the algorithmic complexity is not increased compared to the one of SR model.

The strategy considered in this paper in order to solve the model selection problem can be extended to alternative models: The linear regression could be replaced with an other link or an other distribution could be chosen for the independent variables. If BIC criteria associated to these changes are available, an analogous BIC-like criterion can be derived. Under these modifications, the resulting model should be proved to be identifiable and the construction of the associated algorithm could require deep changes.

A Proof of the criterion consistency theorem

This appendix is devoted to the proof of Theorem 2 addressing the variable selection criterion consistency. This proof is based on the one of the criterion consistency, associated to SR model, given in the Web Supplementary Materials of Maugis et al. (2008) and completely detailed in Maugis (2008).

Proof. According to the expressions (6) and (7), the selected variable partition satisfies $\hat{\mathbf{V}} = \underset{\mathbf{V}\in\mathcal{V}}{\operatorname{argmax}}\ \mathbf{BIC}(\mathbf{V})$ with

$$\mathbf{BIC}(\mathbf{V}) = 2\sum_{i=1}^{n} \ln[f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})] - \Xi_{(\mathbf{V})} \ln(n).$$

Thus

$$P(\hat{\mathbf{V}} = \mathbf{V}_0) = P(\mathbf{BIC}(\mathbf{V}_0) - \mathbf{BIC}(\mathbf{V}) \geq 0, \forall\, \mathbf{V} \in \mathcal{V}). \tag{9}$$

Denoting $\Delta\mathbf{BIC}(\mathbf{V}) = \mathbf{BIC}(\mathbf{V}_0) - \mathbf{BIC}(\mathbf{V})$, we get

$$\begin{aligned}\Delta\mathbf{BIC}(\mathbf{V}) &= 2n\left[\frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}_0})}{h(\mathbf{y}_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}{h(\mathbf{y}_i)}\right\}\right] \\ &\quad + \left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right] \ln(n). \end{aligned} \tag{10}$$

For a variable partition $\mathbf{V} \in \mathcal{V}\backslash\{\mathbf{V}_0\}$, $\mathrm{KL}[h, f(.|\theta^{\star}_{\mathbf{V}})] \neq 0$ since $\theta^{\star}_{\mathbf{V}} \in \Theta'_{\mathbf{V}} \subset \Theta_{\mathbf{V}}$ and according to the model identifiability. Thus, the variable partition set $\mathcal{V}$ can be decomposed into $\mathcal{V} = \{\mathbf{V}_0\} \cup \mathcal{V}_1$ where $\mathcal{V}_1 = \{\mathbf{V} \in \mathcal{V};\ \mathrm{KL}[h, f(.|\theta^{\star}_{\mathbf{V}})] \neq 0\}$. From (9),Theorem 2 is then established if it is proved that

$$\forall\, \mathbf{V} \in \mathcal{V}_1,\ P\left(\Delta\mathbf{BIC}(\mathbf{V}) < 0\right) \underset{n\to\infty}{\to} 0. \tag{11}$$

Let $\mathbf{V} \in \mathcal{V}_1$. Denoting $\mathbb{M}_n(\mathbf{V}) = \frac{1}{n}\sum\limits_{i=1}^{n} \ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}{h(\mathbf{y}_i)}\right\}$ and $M(\mathbf{V}) = -\mathrm{KL}[h, f(.|\theta^{\star}_{\mathbf{V}})]$, from (10) we have

$$\begin{aligned} &P(\Delta\mathbf{BIC}(\mathbf{V}) < 0) = P(2n\{\mathbb{M}_n(\mathbf{V}_0) - \mathbb{M}_n(\mathbf{V})\} + \left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right] \ln(n) < 0) \\ &= P\left(\mathbb{M}_n(\mathbf{V}_0) - M(\mathbf{V}_0) + M(\mathbf{V}_0) - M(\mathbf{V}) + M(\mathbf{V}) - \mathbb{M}_n(\mathbf{V}) + \frac{\left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right] \ln(n)}{2n} < 0\right). \end{aligned}$$

Thus, using the property that for two real random variables A and B and for all $u \in \mathbb{R}$, $P(A + B \leq 0) \leq P(A \leq u) + P(-B > u)$, we get that for all $\epsilon > 0$,

$$\begin{aligned} P(\Delta\mathbf{BIC}(\mathbf{V}) < 0) &\leq P(M(\mathbf{V}_0) - \mathbb{M}_n(\mathbf{V}_0) > \epsilon) + P(\mathbb{M}_n(\mathbf{V}) - M(\mathbf{V}) > \epsilon) \\ &+ P\left(M(\mathbf{V}_0) - M(\mathbf{V}) + \frac{\left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right] \ln(n)}{2n} < 2\epsilon\right). \end{aligned}$$

As in Maugis (2008), it only requires to show that $\forall\, \mathbf{V} \in \mathcal{V}, \mathbb{M}_n(\mathbf{V}) \underset{n\to\infty}{\overset{P}{\to}} M(\mathbf{V})$ in order to prove (11). Thus the proof is finished using the result of the following Lemma 1. □

Lemma 1. *Under assumptions (H1) and (H2),*

$$\forall\, \mathbf{V} \in \mathcal{V},\ \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{h(\mathbf{y}_i)}{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}\right] \underset{n\to\infty}{\overset{P}{\to}} KL[h, f(.|\theta^{\star}_{\mathbf{V}})].$$

Proof. For making easier the reading of this proof, the notation Card(S) is replaced with $\sharp S$ and we recall that all the vectors are implicitly row vectors. Let $\mathbf{V} = (S, R, U, W) \in \mathcal{V}$. As in the proof of Proposition 3.D.1 of Maugis (2008), we want to apply Proposition 2 with the family

$$\mathcal{F}_{(\mathbf{V})} := \{\ln[f(.|\theta)];\ \theta \in \Theta'_{\mathbf{V}}\}$$

in order to obtain

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star_{\mathbf{V}})].$$

Thus we have to prove that (H2) allows to verify the hypotheses of the Proposition 2 and $\mathbb{E}_X[|\ln h(X)|] < \infty$.

Firstly, according to (H2), $\Theta'_{\mathbf{V}}$ is a compact metric space. Moreover, for all $\mathbf{x}$ in $\mathbb{R}^Q$, $\theta_{\mathbf{V}} \in \Theta'_{\mathbf{V}} \mapsto \ln[f(\mathbf{x}|\theta_{\mathbf{V}})]$ is continuous. Let us verify now that there is an envelope function F of $\mathcal{F}_{(\mathbf{V})}$ being h-integrable. Recalling that

$$\ln[f(\mathbf{x}|\theta_{\mathbf{V}})] = \ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] + \ln[f_{\text{reg}}(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega)] + \ln[f_{\text{indep}}(\mathbf{x}^W|\gamma, \tau)],$$

these three terms on the right-hand side are bounded separately. Using the calculus of the proof of Proposition 3.D.1 of Maugis (2008), the two first terms are bounded by

$$-\frac{\sharp S}{2}\ln[2\pi s_{\text{M}}] - \frac{\|\mathbf{x}\|^2 + \eta^2}{s_{\text{m}}} \leq \ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] \leq -\frac{\sharp S}{2}\ln[2\pi s_{\text{m}}] \tag{12}$$

and

$$-\frac{\sharp U}{2}\ln[2\pi s_{\text{M}}] - \frac{\rho^2}{s_{\text{m}}} - \frac{1+\rho^2}{s_{\text{m}}}\|\mathbf{x}\|^2 \leq \ln\left[f_{\text{reg}}(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega)\right] \leq -\frac{\sharp U}{2}\ln[2\pi s_{\text{m}}]. \tag{13}$$

For the third term,

$$\begin{aligned}\ln\left[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)\right] &= \ln\left[|2\pi\tau|^{-1/2}\exp\left(-\frac{1}{2}\|\mathbf{x}^W - \gamma\|^2_{\tau^{-1}}\right)\right] \\ &= -\frac{\sharp W}{2}\ln[2\pi] - \frac{1}{2}\ln[|\tau|] - \frac{1}{2}\|\mathbf{x}^W - \gamma\|^2_{\tau^{-1}}.\end{aligned}$$

Using Lemma 3, the third term can be upper bounded by

$$\ln\left[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)\right] \leq -\frac{\sharp W}{2}\ln[2\pi s_{\text{m}}].$$

According to Lemma 3, $|\tau| \leq s_{\text{M}}^{\sharp W}$ and

$$\begin{aligned}\|\mathbf{x}^W - \gamma\|^2_{\tau^{-1}} &\leq s_{\text{m}}^{-1}\|\mathbf{x}^W - \gamma\|^2 \\ &\leq \frac{2}{s_{\text{m}}}\left(\|\mathbf{x}^W\|^2 + \|\gamma\|^2\right) \\ &\leq \frac{2}{s_{\text{m}}}\left(\|\mathbf{x}\|^2 + \eta^2\right)\end{aligned}$$

because $\gamma \in \mathcal{B}(\eta, \sharp W)$. Then a lower bound of $\ln[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)]$ is

$$\ln[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)] \geq -\frac{\sharp W}{2}\ln[2\pi s_{\text{M}}] - \frac{2(\|\mathbf{x}\|^2 + \eta^2)}{s_{\text{m}}}.$$

Finally the third term is bounded by

$$-\frac{\sharp W}{2}\ln[2\pi s_{\mathrm{M}}] - \frac{2(\|\mathbf{x}\|^2+\eta^2)}{s_{\mathrm{m}}} \leq \ln\left[f_{\mathrm{indep}}(\mathbf{x}^W|\gamma,\tau)\right] \leq -\frac{\sharp W}{2}\ln[2\pi s_{\mathrm{m}}]. \quad (14)$$

Using (12), (13), (14) and $\sharp S + \sharp U + \sharp W = Q$, each function of the family $\mathcal{F}_{(\mathbf{V})}$ is bounded by

$$-\frac{Q}{2}\ln[2\pi s_{\mathrm{M}}] - \frac{2(\|\mathbf{x}\|^2+\eta^2)}{s_{\mathrm{m}}} - \frac{\rho^2}{s_{\mathrm{m}}} - \frac{(1+\rho^2)\|\mathbf{x}\|^2}{s_{\mathrm{m}}} \leq \ln\left[f(\mathbf{x}|\theta_{\mathbf{V}})\right] \leq -\frac{Q}{2}\ln\left[2\pi s_{\mathrm{m}}\right].$$

Thus, for all $\theta_{\mathbf{V}} \in \Theta'_{\mathbf{V}}$ and all $\mathbf{x} \in \mathbb{R}^Q$, $|\ln[f(\mathbf{x}|\theta_{\mathbf{V}})]| \leq C_1(s_{\mathrm{m}}, s_{\mathrm{M}}, Q, \eta, \rho) + C_2(\rho, s_{\mathrm{m}})\|\mathbf{x}\|^2$ defining the envelope function F, where $C_1(s_{\mathrm{m}}, s_{\mathrm{M}}, Q, \eta, \rho)$ and $C_2(\rho, s_{\mathrm{m}})$ are two positive constants. To verify that F is h-integrable, we have to show that $\int \|\mathbf{x}\|^2 h(\mathbf{x})dx < \infty$:

$$\begin{aligned}
\int \|\mathbf{x}\|^2 h(\mathbf{x})d\mathbf{x} &= \int \|\mathbf{x}\|^2 f(\mathbf{x}|\theta^\star_{(S_0,R_0,U_0,W_0)})dx \\
&= \int \|\mathbf{x}\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\mathrm{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) f_{\mathrm{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&= \int \|\mathbf{x}^{S_0}\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{U_0}\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\mathrm{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{W_0}\|^2 f_{\mathrm{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} \\
&\leq \int \|\mathbf{x}^{S_0}\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|a^\star + \mathbf{x}^{R_0}\beta^\star\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|\mathbf{x}^{U_0} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 f_{\mathrm{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\mathrm{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{W_0}\|^2 f_{\mathrm{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0}. \quad (15)
\end{aligned}$$

By a similar study as in Maugis (2008)and Maugis et al. (2008), the three first terms on the right-hand side of Inequality (15) are upper bounded respectively by $2\eta^2 + 2s_{\mathrm{M}}\sharp S_0$, $\rho^2 + \rho^2[2\eta^2 + 2s_{\mathrm{M}}\sharp S_0]$ and $s_{\mathrm{M}}\sharp U_0$. For the fourth term

$$\begin{aligned}
\int \|\mathbf{x}^{W_0}\|^2 f_{\mathrm{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} &= \int \|\mathbf{x}^{W_0}\|^2 \Phi(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} \\
&\leq 2[\|\gamma^\star\|^2 + \mathrm{tr}(\tau^\star)] \\
&\leq 2(\eta^2 + \sharp W_0 s_{\mathrm{M}})
\end{aligned}$$

according to Lemma 4. So turning back to Inequality (15), the integral

$$\int \|\mathbf{x}\|^2 h(\mathbf{x})d\mathbf{x} \leq 4\eta^2 + 2s_{\mathrm{M}}(\sharp S_0 + \sharp W_0) + s_{\mathrm{M}}\sharp U_0 + \rho^2(1 + 2\eta^2 + 2s_{\mathrm{M}}\sharp S_0)$$

and finally F is h-integrable. Since $\ln(h) \in \mathcal{F}_{(S_0,\,R_0,\,U_0,\,W_0)}$, it implies that $\mathbb{E}[|\ln h(X)|] \leq \mathbb{E}[F(X)] < \infty$ and the law of large numbers can be applied to end the proof. □

Proposition 2.
Assume that

1. $(X_1, \ldots, X_n)$ *is a n-sample with unknown density* h.
2. Θ *is a compact metric space.*
3. $\theta \in \Theta \mapsto \ln[f(\boldsymbol{x}|\theta)]$ *is continuous for every* $\boldsymbol{x} \in \mathbb{R}^Q$.
4. F *is an envelope function of* $\mathcal{F} := \{\ln[f(.|\theta)];\ \theta \in \Theta\}$ *which is* h*-integrable.*
5. $\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmax}}\, KL[h, f(.|\theta)]$
6. $\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^n f(X_i|\theta)$.

Then $\frac{1}{n}\sum_{i=1}^{n} \ln\left[f(X_i|\hat{\theta})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star)]$.

This proposition is proved in Maugis (2008).

Lemma 3. *Let* $\Sigma \in \mathcal{D}_r$ *where* $\mathcal{D}_r$ *is defined in (H2). Then*

1. $s_m^r \leq |\Sigma| \leq s_M^r$ *and* $\operatorname{tr}(\Sigma) \leq s_M r$
2. $\forall \mathbf{x} \in \mathbb{R}^r, s_M^{-1}\|\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2_{\Sigma^{-1}} \leq s_m^{-1}\|\mathbf{x}\|^2$

Lemma 4.
Let $\Phi(.|\mu, \Sigma)$ *be the density of the multivariate Gaussian distribution* $\mathcal{N}_r(\mu, \Sigma)$. *Then*

1. $\int \|\mathbf{x}\|^2 \Phi(\mathbf{x}|0, \Sigma) d\mathbf{x} = \operatorname{tr}(\Sigma)$
2. $\int \|\mathbf{x}\|^2 \Phi(\mathbf{x}|\mu, \Sigma) d\mathbf{x} \leq 2\left[\|\mu\|^2 + \operatorname{tr}(\Sigma)\right]$

References

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, 51(2):587–600.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, California.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B.*, 39(1):1–38.

Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166.

Maugis, C. (2008). *Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes*. PhD thesis, Université Paris-Sud 11.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2008). Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*. To appear.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience, New York.

Raftery, A. E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399